Learning by Natural Gradient on Noncompact Matrix-type Pseudo-Riemannian Manifolds

Simone Fiori

Abstract

The present manuscript deals with learning by natural-gradient optimization on non-compact manifolds. In a Riemannian manifold, the calculation of entities such as the closed form of geodesic curves over *non-compact* manifolds might be infeasible. For this reason, it is interesting to study the problem of learning by optimization over non-compact manifolds endowed with pseudo-Riemannian metrics, which may give rise to tractable calculations. A general theory for natural-gradient-based learning on non-compact manifolds as well as specific cases of interest of learning are discussed in this manuscript.

Index Terms

Natural gradient. Learning on non-compact manifolds. Learning by optimization. Geodesic stepping.

I. Introduction

A WIDE class of learning theories for neural systems are formulated in terms of a learning criterion, that measures the performance of a learning system, and of a gradient-based learning algorithm to find the system parameters that optimize its performance.

Natural gradient theory provides a way to learn optimal parameters over an abstract manifold (parameters may be, e.g., the connection weights of a neural system). A seminal paper by Amari [4] introduced and discussed the features of natural gradient in the space $Gl(n)$ of invertible matrices endowed with a Riemannian metric.

The calculation of entities such as the closed form of geodesics over a non-compact manifold endowed with a Riemannian metric might be infeasible. The aim of the present paper is to study the problem of learning by optimization over a non-compact manifold endowed with a pseudo-Riemannian metric, which may give rise to tractable calculations.

An example to motivate the present research is the following. Consider the real symplectic group-manifold which is not flat, as its elements must satisfy a non-linear (quadratic, in fact) constraint. Compute the gradient of a learning criterion to optimize and set-up a gradient-based learning algorithm. Standard Euler discrete stepping to implement the gradient-based learning algorithm (like in standard error back-propagation) is not applicable in this case, because it destroys the symplectic structure. The most natural way to implement the learning steps is via geodesic stepping [6], [7]. However, the form of geodesic arcs for the real symplectic group endowed with a Riemannian metric is not known. The above considerations may be extended to different non-compact manifolds.

In summary, pseudo-Riemannian geometry seems to be the only way to implement effectively a learning algorithm (and to formulate some learning problems) on some non-compact manifolds.

The manuscript is organized as follows. Section II reviews optimization by natural gradient in a Riemannian context and extends such learning framework to the case of pseudo-Riemannian context. Section III describes in details two cases of interest

The author is with Dipartimento di Ingegneria Biomedica, Elettronica e Telecomunicazioni (DiBET), Facoltà di Ingegneria, Università Politecnica delle Marche, Via Brecce Bianche, Ancona I-60131, Italy. (eMail: s.fiori@univpm.it)

and suggests several further cases of non-compact manifolds of interest in the scientific literature. Section IV illustrates the general theory with numerical examples. Section V concludes the paper.

II. Optimization on matrix-type manifolds

In the present section, we briefly review optimization on Riemannian manifolds and discuss the optimization problem over pseudo-Riemannian manifolds. For references on differential geometry, see [18]. In the present manuscript, we are interested in smooth manifolds M that are subsets of the matrix space $\mathbb{R}^{n\times n}$, i.e., matrix-type manifolds.

A. Review of optimization on Riemannian manifolds

Let M be a smooth manifold. The tangent space at $x \in M$ to the manifold is denoted by T_xM.

A Riemannian metric on M is a non-degenerate, smooth, symmetric, positive-definite bilinear map which assigns a real number to pairs of tangent vectors on each tangent space of the manifold M. Let us denote an inner product on M by $\langle .,. \rangle_x : T_xM \times T_xM \to \mathbb{R}$. An inner product must be symmetric and bilinear, therefore, if $u, v, w \in T_xM$ are tangent vectors at a point $x \in M$, then it holds:

- $\langle u, v\rangle_x = \langle v, u\rangle_x$,
- $\langle au + v, w\rangle_x = a\langle u, w\rangle_x + \langle v, w\rangle_x$,

for some $a \in \mathbb{R}$. The requirement that the metric is non-degenerate means that there are no non-zero tangent directions $u \in T_xM$ such that $\langle u, v\rangle_x = 0$ for all $v \in T_xM$. The requirement that the metric is positive-definite means that for any non-zero tangent direction $u \in T_xM$ it holds $\langle u, u\rangle_x > 0$.

In view of gradient-steepest-descent learning by geodesic stepping for criterion optimization, the following quantities are of interest:

- *Closed form of geodesic curves*. Geodesic arcs are the counterparts of straight lines on flat spaces and allow to move on a manifold from a starting point along a given tangent direction. Geodesics also turn a curved space into a metric space as measuring the length of geodesic arcs allows to measure distances between points on a curved space.
- *Closed form of natural gradient*. The natural gradient of a regular function defined over a smooth manifold gives the direction along with the function grows the most around a given point. Natural gradient is thus of prime importance in local optimization.
- *Closed form of geodesic distance*. Geodesic distance is the counterpart of Euclidean distance in Euclidean spaces. The geodesic distance between two points on a manifold is defined as the length of a geodesic curve having those points as endpoints.

When a Riemannian manifold of interest $(M, \langle .,. \rangle_x)$ and a regular learning criterion $f : M \to \mathbb{R}$ are specified, a known optimization rule is gradient steepest descent, that may be expressed by the differential equation:

$$\dot{x} = -\nabla_x f, \tag{1}$$

where symbol ∇_x denotes natural gradient. The flow $x(t)$ associated to such differential equation on M tends toward a local minimum of the criterion function f, in fact:

$$\dot{f} = \langle \nabla_x f, \dot{x}\rangle_x = -\langle \nabla_x f, \nabla_x f\rangle_x \leq 0, \tag{2}$$

with equality holding if and only if $\nabla_x f = 0$, namely, when the flow $x(t)$ approaches a stationary point (local minimum) of the criterion f. Note that:

- the non-positivity of the quantity $\dot{f}$ follows from the positive-definiteness of the metric $\langle .,. \rangle_x$.
- the implication $\langle \nabla_x f, \nabla_x f \rangle_x = 0 \Rightarrow \nabla_x f = 0$ follows from the non-degeneracy of the metric $\langle .,. \rangle_x$.

The basic idea to implement the optimization scheme represented by equation (1) is to replace the continuous-time state $x(t) \in M$ with a discrete-time state $x_k \in M$. This operation requires a numerical integration scheme of the equation (1). In particular, the differential equation on manifold (1) may be solved by the help of the notion of geodesic curve, namely, by the discrete-time optimization stepping rule [13]:

$$x_{k+1} = G_{x_k, -\nabla_{x_k} f}(\eta), \tag{3}$$

where $\eta \in [0\ 1]$ denotes an integration (or learning) step-size (usually $\eta \ll 1$) and where $G_{x,v}(t)$ denotes a geodesic arc departing from the point $x \in M$ with initial direction $v \in T_x M$ and parameter $t \in [0\ 1]$.

If a geodesic arc connecting two points $x_1, x_2 \in M$ exists, it may be found by solving the following variational problem:

$$\delta \int_0^1 \langle \dot{x}(t), \dot{x}(t) \rangle_{x(t)} dt = 0, \tag{4}$$

where symbol δ denotes passing from a point of a curve corresponding to a value of the parameter t to a point of an infinitesimally-close curve corresponding to the same value of the parameter t. Natural parametrization for the geodesic arc was assumed, namely, $\langle \dot{x}(t), \dot{x}(t) \rangle_{x(t)} = \text{constant}$ over the geodesic arc. The variation vanishes at $x(0) = x_1$ and $x(1) = x_2$. It is known that, for a Riemannian manifold, the equation (4) gives rise to the following second-order differential equation:

$$\ddot{x} + \Gamma_x(\dot{x}, \dot{x}) = 0, \tag{5}$$

where Γ_x is the Christoffel operator[1].

The geodesic distance between points $x_1, x_2 \in M$ that are connectible by a geodesic curve $G_{x,v}(t)$ is defined as:

$$d(x_1, x_2) \stackrel{\text{def}}{=} \int_0^1 \langle \dot{G}_{x,v}(t), \dot{G}_{x,v}(t) \rangle^{\frac{1}{2}}_{G_{x,v}(t)} dt \tag{6}$$

The natural gradient of a regular function $f : M \to \mathbb{R}$ may be defined in several ways. Let the manifold M and the tangent spaces $T_x M$, $x \in M$, be defined implicitly as:

$$M \stackrel{\text{def}}{=} \{x \in \mathbb{R}^{n \times n} | \varphi(x) = 0\}, \tag{7}$$

$$T_x M \stackrel{\text{def}}{=} \{v \in \mathbb{R}^{n \times n} | \varphi'_x(v) = 0\}, \tag{8}$$

where φ is a non-linear matrix-to-matrix function and φ'_x is a linear matrix-to-matrix operator in the variable v. By generalizing the construction of Amari [4], here we adopt the following definition: The natural gradient $\nabla_x f$ is the minimizer of the function:

$$F_x(v) \stackrel{\text{def}}{=} -\operatorname{tr}(v^T \partial_x f) + \frac{\lambda_x}{2}(\langle v, v \rangle_x - c_x) + \operatorname{tr}(\omega_x \varphi'_x(v)), \tag{9}$$

where $\partial_x f$ denotes Euclidean gradient, $\lambda_x \in \mathbb{R}$ is a Lagrange multiplier for the constraint $\langle v, v \rangle_x = c_x > 0$ and $\omega_x \in \Omega$ is a Lagrange-multiplier matrix for the constraint $\varphi'_x(v) = 0$ for some appropriate matrix-space Ω. The quantities F_x, λ_x, c_x and ω_x may depend on the point x and the notation reflects such dependency. Symbol tr denotes matrix trace. A well-known property of the natural gradient in Riemannian geometry, that the definition of functional (9) is consistent with, is that: *The natural gradient of a function points toward the direction along with the function locally increases the most.*

[1]On a Riemannian manifold, the Christoffel operator is completely determined by the metric and may be arbitrarily complicated even for simple metrics such as the metric induced by the Euclidean one.

In the case that the manifold M is a matrix Lie group with Lie algebra $\mathfrak{m}$, a way to define a metric is by exploiting the Euclidean metric in the Lie algebra:

$$\langle u, u\rangle_{\mathfrak{m}} \stackrel{\text{def}}{=} \frac{1}{2}\text{tr}(u^T u),\ \ u \in \mathfrak{m}, \tag{10}$$

to be translated to each tangent space by left-translation:

$$\langle v, v\rangle_x = \langle x^{-1} \cdot v, x^{-1} \cdot v\rangle_{\mathfrak{m}}, \tag{11}$$

where x^{-1} denotes the inverse of the element $x \in M$ according to the algebraic group structure of the Lie group M and $\cdot$ denotes group multiplication. Right-translation might be of course used, instead. This is the approach that was used by Amari [4] and that gives rise to the natural metric for the Lie group $Gl(n)$:

$$\langle v, v\rangle_x = \frac{1}{2}\text{tr}(v^T (xx^T)^{-1} v), \tag{12}$$

that corresponds to the natural gradient $\nabla_x f = (xx^T)\partial_x f$.

Although such induced metric looks very natural, in some settings of interest it might give rise to geodesic equations whose closed-form solutions are unknown. Conversely, non-Riemannian (namely, pseudo-Riemannian) metrics might give rise to tractable geodesic equations and geodesic distances.

B. Optimization on pseudo-Riemannian manifolds

A pseudo-Riemannian manifold is a differentiable manifold equipped with a smooth, symmetric metric that may possess positive as well as negative eigenvalues, but *not null eigenvalues*. Such a metric is called a pseudo-Riemannian metric and its values can be positive, zero or even negative[2].

Examples of pseudo-Riemannian metrics are the Minkowski metric and, more generally, the Lorentz metric, that provide a ground for the general theory of relativity. The flat Minkowski space $(\mathbb{R}^4, \langle ., .\rangle_x)$ has metric:

$$\langle v, v\rangle_x \stackrel{\text{def}}{=} v^T K v,\ \ K \stackrel{\text{def}}{=} \text{diag}(+1, +1, +1, -1).$$

Clearly, there exist non-zero directions $v \in \mathbb{R}^4$ such that $v^T K v > 0$ (space-like direction), $v^T K v < 0$ (time-like direction) and even $v^T K v = 0$ (light-like direction). However, note that the linear map $v \mapsto \frac{\partial}{\partial v}\langle v, v\rangle_x$ is invertible for every element x on the manifold.

An example of pseudo-Riemannian metric for $M = Gl(n)$ is [15]:

$$\langle v, v\rangle_x \stackrel{\text{def}}{=} \text{tr}((x^{-1} v)^2). \tag{13}$$

Apparently, there might exist tangent vectors $v \in T_x M$ such that $\langle v, v\rangle_x < 0$ as well as non-zero vectors for which $\langle v, v\rangle_x = 0$.

A fundamental idea to deal with pseudo-Riemannian metrics is that any tangent space $T_x M$ of a pseudo-Riemannian manifold may be decomposed according to its pseudo-Riemannian metric as:

$$\begin{cases} T_x M & = & T_x^+ M \cup T_x^0 M \cup T_x^- M, \\ T_x^+ M & \stackrel{\text{def}}{=} & \{v \in T_x M | \langle v, v\rangle_x > 0\} \cup \{0\}, \\ T_x^0 M & \stackrel{\text{def}}{=} & \{v \in T_x M | \langle v, v\rangle_x = 0, v \neq 0\}, \\ T_x^- M & \stackrel{\text{def}}{=} & \{v \in T_x M | \langle v, v\rangle_x < 0\}. \end{cases} \tag{14}$$

[2]Pseudo-Riemannian manifolds endowed with a metric that presents null eigenvalues may be considered too and are termed 'singular pseudo-Riemannian manifolds' [16]. In the present manuscript, we do not treat the case of singular pseudo-Riemannian manifolds.

Such decomposition links the notion of pseudo-Riemannian manifold to the one of Riemannian manifold: Intuitively, the part T_x^+M is one where the pseudo-Riemannian metric behaves as a Riemannian metric, while the part T_x^-M is one where the metric behaves as an anti-Riemannian metric.

The question is how to define the notion of geodesic curves, geodesic distances and natural gradient, in presence of a pseudo-Riemannian metric, for optimization purpose.

The geodesic curves on a pseudo-Riemannian manifold $(M, \langle ., . \rangle_x)$ may be defined again by the variational principle (4). Now, by the stationarity of the integrand over a solution, it holds:

$$\langle \dot{G}_{x,v}(t), \dot{G}_{x,v}(t) \rangle_{G_{x,v}(t)} = k_{x,v}^2, \ \forall t \in [0\ 1], \tag{15}$$

with $k_{x,v}^2 \in \mathbb{R}$ independent of the parameter t. The quantity $k_{x,v}^2$ is actually the squared distance between the endpoints $G_{x,v}(0) = x_1$ and $G_{x,v}(1) = x_2$, namely:

$$k_{x,v}^2 = d^2(x_1, x_2). \tag{16}$$

This further implies that:

$$d^2(G_{x,v}(t_1), G_{x,v}(t_2)) = (t_2 - t_1)^2 k_{x,v}^2, \tag{17}$$

for every pair $t_1, t_2 \in [0\ 1]$. Therefore, we have the following consistency result for geodesics on a pseudo-Riemannian manifold:

- If the squared distance between endpoints is positive, namely $k_{x,v}^2 > 0$, then the squared distance between any pair of distinct points on a geodesic arc is positive, namely $d^2(G_{x,v}(t_1), G_{x,v}(t_2)) > 0$ for every pair $t_1, t_2 \in [0\ 1]$ with $t_2 \neq t_1$.
- If the squared distance between endpoints is zero, namely $k_{x,v} = 0$, then the squared distance between any pair of points on a geodesic arc is zero.
- If the squared distance between endpoints is negative, namely $k_{x,v}^2 < 0$, then the squared distance between any pair of distinct points on a geodesic arc is negative, namely $d^2(G_{x,v}(t_1), G_{x,v}(t_2)) < 0$ for every pair $t_1, t_2 \in [0\ 1]$ with $t_2 \neq t_1$.
- The squared distance between two coincident points is zero.

The above discussion suggests that the space L of geodesics over a pseudo-Riemannian manifold M may be decomposed as:

$$L = L^+ \cup L^0 \cup L^-, \tag{18}$$

where L^+ denotes the set of 'space-like' geodesics $(k_{x,v}^2 > 0)$, L^0 denotes the set of 'light-like' geodesics $(k_{x,v} = 0)$ and L^- denotes the set of 'time-like' geodesics $(k_{x,v}^2 < 0)$ [14].

The natural gradient on a pseudo-Riemannian manifold expressed as in (7)-(8) may be defined as an extension of the principle (9).

First, let us reformulate the function F_x in a more convenient way. The term $\mathrm{tr}(\omega_x \varphi_x'(v)))$ in (9) is linear in the argument v, therefore it may be rewritten as $\mathrm{tr}(v^T \alpha_x(\omega_x))$ for some function α_x. The function F_x may thus be rewritten as:

$$F_x(v) = \frac{\lambda_x}{2} \langle v, v \rangle_x + \mathrm{tr}(v^T(\alpha_x(\omega_x) - \partial_x f)) - \frac{\lambda_x}{2} c_x, \tag{19}$$

where $c_x \neq 0$ and may be negative as well as positive. Let us define the map:

$$\psi_x(v) \stackrel{\mathrm{def}}{=} \frac{1}{2} \frac{\partial}{\partial v} \langle v, v \rangle_x, \tag{20}$$

that, for a pseudo-Riemannian metric, is invertible for every $x \in M$. The stationarity conditions for the Lagrangean (19) may be written:

$$\begin{cases} \lambda_x \psi_x(v) + \alpha_x(\omega_x) - \partial_x f = 0, \\ \lambda_x \langle v, v \rangle_x = c_x, \\ \varphi'_x(v) = 0. \end{cases} \tag{21}$$

The above set of equations may be rearranged in the following explicit form:

$$\lambda_x v = \psi_x^{-1}(\partial_x f - \alpha_x(\omega_x)), \tag{22}$$

$$\lambda_x^2 c_x = \langle \psi_x^{-1}(\partial_x f - \alpha_x(\omega_x)), \psi_x^{-1}(\partial_x f - \alpha_x(\omega_x)) \rangle_x, \tag{23}$$

$$\varphi'_x(\psi_x^{-1}(\partial_x f - \alpha_x(\omega_x))) = 0, \ \omega_x \in \Omega. \tag{24}$$

The condition (23) serves to determine the appropriate squared value of the Lagrange multiplier λ_x. As the quantity c_x is arbitrary (albeit small), we may assume:

$$\lambda_x^2 = 1. \tag{25}$$

The equation (22) gives the expression of the natural gradient. Note that the squared norm of the natural gradient $\langle v, v \rangle_x$ is independent of the sign of the Lagrange multiplier $\lambda_x = \pm 1$. Three cases should be distinguished among:

- Case $v \in T_x^+ M$ (Riemannian-like): In this case, the parabolic function $F_x(v)$ in (19) has convexity up for $\lambda_x = +1$, hence the natural gradient is:

$$\nabla_x f = \psi_x^{-1}(\partial_x f - \bar{\alpha}_x), \ \text{for } \psi_x^{-1}(\partial_x f - \bar{\alpha}_x) \in T_x^+ M. \tag{26}$$

- Case $v \in T_x^- M$ (anti-Riemannian-like): In this case, the parabolic function $F_x(v)$ in (19) has convexity up for $\lambda_x = -1$, hence the natural gradient is:

$$\nabla_x f = -\psi_x^{-1}(\partial_x f - \bar{\alpha}_x), \ \text{for } \psi_x^{-1}(\partial_x f - \bar{\alpha}_x) \in T_x^- M. \tag{27}$$

- Case $v \in T_x^0 M$: In this event, the function $F_x(v)$ is linear in v and does not admit any minimizer. Therefore, the natural gradient is undefined:

$$\nabla_x f \text{ not defined for } \psi_x^{-1}(\partial_x f - \bar{\alpha}_x) \in T_x^0 M. \tag{28}$$

The first two cases above may be summarized as:

$$\nabla_x f = \pm \psi_x^{-1}(\partial_x f - \bar{\alpha}_x), \ \text{for } \partial_x f - \bar{\alpha}_x \in \psi_x(T_x^{\pm} M), \tag{29}$$

with clear meaning of symbols.

C. Pseudo-Riemannian metric compatibility property and its implications

A brief investigation on the counterpart of the metric compatibility of the natural gradient extended to the pseudo-Riemannian manifolds is in order. As this is a theoretical result that does not need implementation, it is convenient to work in local coordinates: The local-coordinates-counterparts of the entities defined in the previous subsections will be denoted by a 'hat' and it will be assumed that they belong to Euclidean spaces of appropriate dimensions.

In local coordinates, it holds:

$$\langle \hat{v}, \hat{v} \rangle_{\hat{x}} = \hat{v}^T K_{\hat{x}} \hat{v}, \ K_{\hat{x}}^T = K_{\hat{x}}, \tag{30}$$

$$F_{\hat{x}}(\hat{v}) = -\hat{v}^T \hat{\partial}_{\hat{x}} f + \tfrac{1}{2} \lambda_{\hat{x}} (\langle \hat{v}, \hat{v} \rangle_{\hat{x}} - c_{\hat{x}}). \tag{31}$$

Now, away from the light-like part of the tangent bundle, the minimization of the above functional gives the pseudo-Riemannian counterpart of a well-known result for Riemannian metrics, namely: $\hat{\nabla}_{\hat{x}} f = \pm K_{\hat{x}}^{-1} \hat{\partial}_{\hat{x}} f$. Hence, the following result holds: $\langle \hat{\nabla}_{\hat{x}} f, u \rangle_{\hat{x}} = \pm u^T \hat{\partial}_{\hat{x}} f$, for every $u \in T_{\hat{x}} M$. This is referred to as metric compatibility property of the natural gradient. In intrinsic coordinates, for matrix-type manifolds, the above result may be recast as:

$$\langle u, \nabla_x f \rangle_x = \pm \mathrm{tr}(u^T \partial_x f), \ \forall u \in T_x M. \tag{32}$$

The above metric compatibility condition has an important application in optimization. In fact, the differential equation $\dot{x} = -\nabla_x f$ induces the following dynamics:

$$\begin{aligned}
\frac{d}{dt} f(x(t)) &= \mathrm{tr}(\dot{x}^T(t) \partial_{x(t)} f(x(t))) \\
&= \begin{cases} -\langle \nabla_{x(t)} f(x(t)), \nabla_{x(t)} f(x(t)) \rangle_{x(t)} & \text{for } \nabla_{x(t)} f(x(t)) \in T^+_{x(t)} M, \\ \langle \nabla_{x(t)} f(x(t)), \nabla_{x(t)} f(x(t)) \rangle_{x(t)} & \text{for } \nabla_{x(t)} f(x(t)) \in T^-_{x(t)} M, \\ \text{Undefined} & \text{for } \nabla_{x(t)} f(x(t)) \in T^0_{x(t)} M, \end{cases} \\
&= \begin{cases} -|\langle \nabla_{x(t)} f(x(t)), \nabla_{x(t)} f(x(t)) \rangle_{x(t)}| & \text{for } \nabla_{x(t)} f(x(t)) \notin T^0_{x(t)} M, \\ \text{Undefined} & \text{for } \nabla_{x(t)} f(x(t)) \in T^0_{x(t)} M. \end{cases}
\end{aligned}$$

The above equations confirm that the optimization differential equation $\dot{x} = -\nabla_x f$ actually 'seeks' for a local minimum of the function f except that in a set of impasse points related to the light-like part of the tangent bundle.

On the other hand, the above equations suggest a way to define a pseudo-natural gradient even in the singularity parts $T^0_x M$. In fact, if we define the natural gradient to be:

$$\bar{\nabla}_x f = \begin{cases} \psi_x^{-1}(\partial_x f - \bar{\alpha}_x), & \text{for } \partial_x f - \bar{\alpha}_x \in \psi_x(T^+_x M \cup T^0_x M), \\ -\psi_x^{-1}(\partial_x f - \bar{\alpha}_x), & \text{for } \partial_x f - \bar{\alpha}_x \in \psi_x(T^-_x M), \end{cases} \tag{33}$$

then the natural gradient is always defined and moreover the differential equation $\dot{x} = -\bar{\nabla}_x f$ implies:

$$\dot{f} = \begin{cases} -|\langle \bar{\nabla}_x f, \bar{\nabla}_x f \rangle_x|, & \text{for } \bar{\nabla}_x f \notin T^0_x M, \\ 0, & \text{for } \bar{\nabla}_x f \in T^0_x M. \end{cases} \tag{34}$$

Namely, whenever $\bar{\nabla}_x f \in T^0_x M$ the optimization differential equation does not change the value of the criterion function even if the optimization differential equation 'moves' infinitesimally the current point along the direction $\bar{\nabla}_x f$.

All in one, the dynamics (34) suggests that pseudo-Riemannian-gradient-based learning rules share the same convergence properties of natural-gradient-based learning rules, except for a null-measure set of impasse points.

III. Discussion of cases of interest

In the present section, we discuss the details about optimization on two non-compact manifolds of interest. We also mention several cases of interest in science.

A. Optimization on the real general linear group as a pseudo-Riemannian manifold

In the present section, we apply the definitions of the previous section to the real general linear group endowed with the metric (13). The general linear group is defined as:

$$Gl(n) \stackrel{\mathrm{def}}{=} \{x \in \mathbb{R}^{n \times n} | \det(x) \neq 0\}. \tag{35}$$

The tangent spaces of the real general linear group exhibit the structure:

$$T_x Gl(n) = \mathbb{R}^{n\times n}. \tag{36}$$

If the group-manifold $Gl(n)$ is endowed with the metric (13), the geodesic equation writes:

$$\delta \int_0^1 \mathrm{tr}(x^{-1}\dot{x}x^{-1}\dot{x})dt = 0,\ \delta x \in T_x Gl(n), \tag{37}$$

where the natural parametrization is assumed. The left-hand side rewrites:

$$\delta \int_0^1 \mathrm{tr}(x^{-1}\dot{x}x^{-1}\dot{x})dt = 2\int_0^1 \mathrm{tr}(\delta(x^{-1})\dot{x}x^{-1}\dot{x})dt + 2\int_0^1 \mathrm{tr}(\delta\dot{x}x^{-1}\dot{x}x^{-1})dt. \tag{38}$$

In the first integral of the right-hand side, it holds $\delta(x^{-1}) = -x^{-1}\delta x x^{-1}$, while in the second integral of the right-hand side it holds $\delta\dot{x} = \frac{d}{dt}\delta x$. Now, integration by parts gives:

$$\int_0^1 \mathrm{tr}(\frac{d\delta x}{dt}x^{-1}\dot{x}x^{-1})dt = -\int_0^1 \mathrm{tr}(\delta x \frac{d}{dt}(x^{-1}\dot{x}x^{-1}))dt, \tag{39}$$

because the variation δx vanishes at endpoints of the geodesic (namely at $t = 0$ and $t = 1$). The geodesic equation becomes then:

$$\int_0^1 \mathrm{tr}(\delta x(x^{-1}\dot{x}x^{-1}\dot{x}x^{-1} + \frac{d}{dt}(x^{-1}\dot{x}x^{-1})))dt = 0. \tag{40}$$

As the variation $\delta x \in \mathbb{R}^{n\times n}$ is arbitrary, the above equation holds if and only if:

$$\frac{d}{dt}(x^{-1}\dot{x}x^{-1}) + x^{-1}\dot{x}x^{-1}\dot{x}x^{-1} = 0. \tag{41}$$

By using the fact that $\frac{d}{dt}x^{-1} = -x^{-1}\dot{x}x^{-1}$, the above equation may be reduced to the simpler form:

$$\ddot{x} - \dot{x}x^{-1}\dot{x} = 0, \tag{42}$$

which is in the form (5), in fact. Noticeably, the equation (42) admits the closed form solution:

$$G_{x,v}(t) = x\exp(tx^{-1}v), \tag{43}$$

as it may be readily verified by substitution.

Let us consider two points $x_1, x_2 \in Gl(n)$ and assume that they may be joined by a geodesic arc $G_{x,v}(t)$, namely, that there exist $x \in Gl(n)$ and $v \in T_x Gl(n)$ such that $G_{x,v}(0) = x_1$ and $G_{x,v}(1) = x_2$. The two geodesic parameters may be determined easily and have values $x = x_1$, $v = x_1\log(x_1^{-1}x_2)$. Also, it holds $\dot{G}_{x,v}(t) = v\exp(tx^{-1}v)$. Therefore:

$$\begin{aligned} d(x_1,x_2) &= \int_0^1 \mathrm{tr}^{\frac{1}{2}}(((G_{x,v}(t))^{-1}\dot{G}_{x,v}(t))^2)dt \\ d^2(x_1,x_2) &= k_{x,v}^2 = \mathrm{tr}((x^{-1}v)^2) \\ &= \mathrm{tr}(\log^2(x_1^{-1}x_2)). \end{aligned} \tag{44}$$

Apparently, the squared geodesic distance between two non-coincident points may be positive, zero, or even negative, according to the pseudo-Riemannian nature of the space $Gl(n)$ endowed with the pseudo-Riemannian metric (13).

The matrix logarithm may be defined by the series:

$$\log x = -\sum_{k=1}^{\infty}\frac{(e_n - x)^k}{k}, \tag{45}$$

where symbol e_n denotes the identity matrix of size $n \times n$. The series converges for $\|x - e_n\| < 1$. According to the optimization principle (9), in the present case the natural gradient of a regular function $f : Gl(n) \to \mathbb{R}$ may be defined as the minimizer of the functional:

$$F_x(v) = -\mathrm{tr}(v^T \partial_x f) + \frac{\lambda_x}{2}(\mathrm{tr}(x^{-1} v x^{-1} v) - c_x). \tag{46}$$

In this case, the operator ψ_x has expression:

$$\psi_x(v) = x^{-T} v^T x^{-T}, \tag{47}$$

which is invertible for any $x \in Gl(n)$, in fact $\psi_x^{-1}(u) = x u^T x$. Therefore, it holds:

$$\psi_x^{-1}(\partial_x f) = x \partial_x^T f x, \tag{48}$$

$$\langle \psi_x^{-1}(\partial_x f), \psi_x^{-1}(\partial_x f) \rangle_x = \mathrm{tr}((x^T \partial_x f)^2). \tag{49}$$

Hence, the natural gradient of the function f may be defined as follows:

$$\nabla_x f \stackrel{\mathrm{def}}{=} \begin{cases} x \partial_x^T f x, & \text{for } \mathrm{tr}((x^T \partial_x f)^2) > 0, \\ -x \partial_x^T f x, & \text{for } \mathrm{tr}((x^T \partial_x f)^2) < 0, \\ \text{Undefined}, & \text{for } \mathrm{tr}((x^T \partial_x f)^2) = 0. \end{cases} \tag{50}$$

It is easy to study the behavior of the equation:

$$\frac{d}{dt} x(t) = -\nabla_{x(t)} f(x(t)). \tag{51}$$

under the action of the natural gradient (50). As it holds:

$$\begin{aligned} \frac{df}{dt} &= \mathrm{tr}(\dot{x}^T \partial_x f) = \begin{cases} -\mathrm{tr}((x^T \partial_x f)^2), & \text{for } \mathrm{tr}((x^T \partial_x f)^2) > 0, \\ \mathrm{tr}((x^T \partial_x f)^2), & \text{for } \mathrm{tr}((x^T \partial_x f)^2) < 0, \end{cases} \\ &= -|\mathrm{tr}((x^T \partial_x f)^2)|, \end{aligned}$$

then the gradient-based optimization system (51) seeks for a local minimum of the criterion f, except that for a null-measure set of impasse points for which $\mathrm{tr}((x^T \partial_x f)^2) = 0$.

As an example of calculation, let us fix an element $g \in Gl(n)$ and compute the natural pseudo-Riemannian gradient of the map $x \mapsto d^2(x, g)$. The natural gradient of the function $d^2(x, g)$ may be computed as follows. First, note that:

$$\partial_x^T d^2(g, x) = 2 \log(g^{-1} x) x^{-1}, \tag{52}$$

and, therefore:

$$\mathrm{tr}((x^T \partial_x d^2(g, x))^2) = 4 \mathrm{tr}(\log^2(g^{-1} x)). \tag{53}$$

Plugging equations (52) and (53) into equation (50), the following expression is readily obtained:

$$\bar{\nabla}_x d^2(g, x) = \begin{cases} 2x \log(g^{-1} x), & \text{for } \mathrm{tr}(\log^2(g^{-1} x)) \geq 0, \\ -2x \log(g^{-1} x), & \text{for } \mathrm{tr}(\log^2(g^{-1} x)) < 0. \end{cases} \tag{54}$$

B. Optimization on the real symplectic group as a pseudo-Riemannian manifold

In the present section, we apply the definitions of the Section II to the real symplectic group endowed with the metric (13). The real symplectic group is defined as:

$$Sp(2n) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^{2n\times 2n} | x^T q x = q\}, \tag{55}$$

where q is termed fundamental skew-symmetric matrix and is defined as:

$$q \stackrel{\text{def}}{=} \begin{bmatrix} 0_n & e_n \\ -e_n & 0_n \end{bmatrix}, \tag{56}$$

where symbols 0_n and e_n denote the zero matrix and the identity matrix of size $n \times n$, respectively. The tangent spaces of the real symplectic group possess the structures:

$$T_x Sp(2n) = \{v \in \mathbb{R}^{2n\times 2n} | v^T q x + x^T q v = 0\}. \tag{57}$$

The following definitions and property are of use in the theory of symplectic matrices:

$$\mathfrak{s}^+(2n) \stackrel{\text{def}}{=} \{\sigma \in \mathbb{R}^{2n\times 2n} | \sigma^T = \sigma\}, \tag{58}$$

$$\mathfrak{so}(2n) \stackrel{\text{def}}{=} \{\omega \in \mathbb{R}^{2n\times 2n} | \omega^T + \omega = 0\}, \tag{59}$$

$$q^T = q^{-1} = -q,\ q^2 = -e_{2n}, \tag{60}$$

$$T_x Sp(2n) = \{xq\sigma | \sigma \in \mathfrak{s}^+(2n)\}. \tag{61}$$

A study of the geodesic arcs on the real symplectic groups $Sp(2n)$ was recently published by Bloch *et al.* [5]. In such contribution, the space $Sp(2n)$ is regarded as a Riemannian manifold, which results in a equation for the geodesic arcs (5) whose closed-form solution is unknown.

If the Lie group $Sp(2n)$ is endowed with the metric (13), then the geodesic equation writes:

$$\delta \int_0^1 \mathrm{tr}(x^{-1}\dot{x}x^{-1}\dot{x})dt = 0,\ \delta x \in T_x Sp(2n). \tag{62}$$

where the natural parametrization is assumed. By computing the variations, integrating by parts and recalling that the variation vanishes at endpoints, we found that the above principle implies:

$$\int_0^1 \mathrm{tr}(\delta x(x^{-1}\ddot{x}x^{-1} - x^{-1}\dot{x}x^{-1}\dot{x}x^{-1}))dt = 0. \tag{63}$$

The variation $\delta x \in T_x Sp(2n)$ is arbitrary. Now the equation $\mathrm{tr}(\nu^T \delta x) = 0$, with $\delta x \in T_x Sp(2n)$, implies that $\nu^T = \omega q x^{-1}$ with $\omega \in \mathfrak{so}(2n)$, therefore the above equation is satisfied if and only if:

$$x^{-1}\ddot{x}x^{-1} - x^{-1}\dot{x}x^{-1}\dot{x}x^{-1} = \omega q x^{-1},\ \omega \in \mathfrak{so}(2n),$$

or, equivalently,

$$\ddot{x} - \dot{x}x^{-1}\dot{x} = x\omega q, \tag{64}$$

for some $\omega \in \mathfrak{so}(2n)$. In order to determine the value of matrix ω, it is worth observing that:

$$x^T q x - q = 0 \Rightarrow \ddot{x}^T q x + 2\dot{x}^T q \dot{x} + x^T \ddot{x} = 0.$$

Substituting the expression $\ddot{x} = \dot{x}x^{-1}\dot{x} + x\omega q$ into the above equation gives the condition:

$$q\omega q = 0. \tag{65}$$

Hence, $\omega = 0$ and the geodesic equation coincides to equation (42). Therefore, the geodesic curve and the geodesic squared distance have once again the following expressions:

$$G_{x,v}(t) = x\exp(tx^{-1}v), \tag{66}$$

$$d^2(x_1, x_2) = \mathrm{tr}(\log^2(x_1^{-1}x_2)). \tag{67}$$

According to the optimization principle (9), in the present case the natural gradient of a regular function $f : Sp(2n) \to \mathbb{R}$ coincides with the minimizer of functional:

$$F_x(v) = -\mathrm{tr}(v^T\partial_x f) + \frac{\lambda_x}{2}(\mathrm{tr}(x^{-1}vx^{-1}v) - c_x) + \frac{1}{2}\mathrm{tr}(\omega_x(v^Tqx + x^Tqv)), \tag{68}$$

with $\lambda_x \in \mathbb{R}$ and $\omega_x \in \Omega$, where, for symmetry reasons, it is necessary to set $\Omega = \mathfrak{so}(2n)$. As $\lambda_x = \pm 1$, the natural gradient may be found by the following pair of conditions:

$$\begin{cases} -\partial_x f + \lambda_x x^{-T}v^Tx^{-T} + qx\omega_x = 0 \\ v^Tqx + x^Tqv = 0. \end{cases} \tag{69}$$

Solving the above system yields:

$$\omega_x = -\frac{1}{2}(\partial_x^T f xq + x^{-1}q\partial_x f),$$
$$\lambda_x v = \frac{1}{2}q(\partial_x f)q + \frac{1}{2}x(\partial_x f)^T x.$$

On the basis of the following definitions:

$$\tilde{v}_x \stackrel{\mathrm{def}}{=} \frac{1}{2}(q\partial_x f q + x\partial_x^T f x), \tag{70}$$

$$\|\tilde{v}_x\|_x^2 \stackrel{\mathrm{def}}{=} \mathrm{tr}((x^{-1}\tilde{v}_x)^2), \tag{71}$$

the natural gradient $\bar{\nabla}_x f$ may be expressed as:

$$\bar{\nabla}_x f = \tilde{v}_x \mathrm{sign}(\|\tilde{v}_x\|_x^2), \ \|\tilde{v}_x\|_x \neq 0. \tag{72}$$

C. Further non-compact manifolds of interest

Some further cases of interest of learning by optimization on non-compact manifolds are as follows:

- Optimization on the special linear group $Sl(n)$, with application to approximate joint diagonalization [3]. Simultaneous diagonalization of a set of matrices is part of many algorithms. The early methods developed for simultaneous diagonalization restricted the joint diagonalizer to belong to the compact Lie group of orthogonal matrices. Nonetheless, non-orthogonal joint diagonalization is very appealing in signal-processing and machine-learning applications [3], [19].
- Optimization on the oblique manifold, with application to approximate joint diagonalization [1]. Several blind source separation algorithms compute a separating matrix that approximately diagonalizes a collection of covariance matrices. Since the magnitude of the sources is unknown, there is a fundamental indeterminacy on the norm of the rows of the separating matrix. Such indeterminacy can be taken into account by restricting the separating matrix to the oblique manifold.
- Subspace analysis (e.g., invariant subspace computation and computation of the mean of subspaces) may be cast as optimization on the Grassmann manifold $Gr(n, p)$ and the *non-compact* Stiefel manifold $ST(n, p)$ [2].

- Representations in terms of elements of the space $\Theta^+(n)$ of the upper triangular matrices with positive diagonal elements, and of the space $S\Theta^+(n)$ of the upper triangular matrices with positive diagonal elements whose product is unitary, find applications in the representation of shapes in pattern recognition [17].

It is also worth mentioning the non-compact manifold formed by symmetric positive-definite matrices $S^+(n)$, which finds a large number of applications [8]. Such a group-manifold is one of the few encountered exceptions of non-compact manifolds that admit closed-form expressions of geodesic arcs and geodesic distances when endowed with a Riemannian metric. Noticeably, however, the metric (13) coincides to the natural Riemannian metric for the space $S^+(n)$.

IV. Examples

The present section illustrates some examples. Section IV-A shows an example of calculation of pseudo-natural-gradient in the real plane and a comparison with the Euclidean gradient. Sections IV-B presents numerical experiments of learning by optimization on the general linear group and on the real symplectic group.

A. The real plane endowed with a Lorentz metric

The purpose of the example discussed within the present subsection is to illustrate the definitions given in section II and the calculations encountered in the determination of pseudo-Riemannian gradient and pseudo-Riemannian geodesic arcs. Also, thanks to the low dimensionality of the spaces of interest, in this example it is possible to render them graphically (unfortunately, this option is not amenable for other manifolds of interest which are, generally, of much higher dimension).

In the present subsection, standard notation for vectors in $\mathbb{R}^2$ is made use of, namely, a point of the real plane is denoted by $(x, y) \in \mathbb{R}^2$ and thus $(v_x, v_y) \in T_{(x,y)}\mathbb{R}^2$.

Consider the pseudo-Riemannian (Lorentz) manifold $(\mathbb{R}^2, \langle ., . \rangle_{(x,y)})$ with:

$$\langle (v_x, v_y), (v_x, v_y) \rangle_{(x,y)} \stackrel{\text{def}}{=} \begin{bmatrix} v_x \\ v_y \end{bmatrix}^T \begin{bmatrix} +1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} v_x \\ v_y \end{bmatrix} = v_x^2 - v_y^2, \tag{73}$$

and let us define the function $f : \mathbb{R}^2 \to \mathbb{R}$ to minimize as:

$$f(x, y) = x^2 + 2y^2. \tag{74}$$

Apparently, the minimum of the function f above locates in $(0, 0)$.

The tangent spaces to the pseudo-Riemannian manifold $(\mathbb{R}^2, \langle ., . \rangle_{(x,y)})$ may be decomposed as follows:

$$T_{(x,y)}\mathbb{R}^2 = T^+_{(x,y)}\mathbb{R}^2 \cup T^0_{(x,y)}\mathbb{R}^2 \cup T^-_{(x,y)}\mathbb{R}^2, \tag{75}$$

where:

$$\begin{cases} T^+_{(x,y)}\mathbb{R}^2 = \{(v_x, v_y) \in \mathbb{R}^2 | v_x^2 > v_y^2\} \cup \{(0, 0)\}, \\ T^0_{(x,y)}\mathbb{R}^2 = \{(1, 1)v | v \in \mathbb{R} - \{0\}\} \cup \{(1, -1)v | v \in \mathbb{R} - \{0\}\}, \\ T^-_{(x,y)}\mathbb{R}^2 = \{(v_x, v_y) \in \mathbb{R}^2 | v_x^2 < v_y^2\}. \end{cases} \tag{76}$$

The structure of the decomposed tangent space $T_{(x,y)}\mathbb{R}^2$ is shown in the Figure 1.

The functional to optimize in order to define the natural gradient of function f is:

$$F_{(x,y)}(v_x, v_y) = -(v_x \partial_x f + v_y \partial_y f) + \frac{\lambda_{(x,y)}}{2}(v_x^2 - v_y^2 - c_{(x,y)}). \tag{77}$$

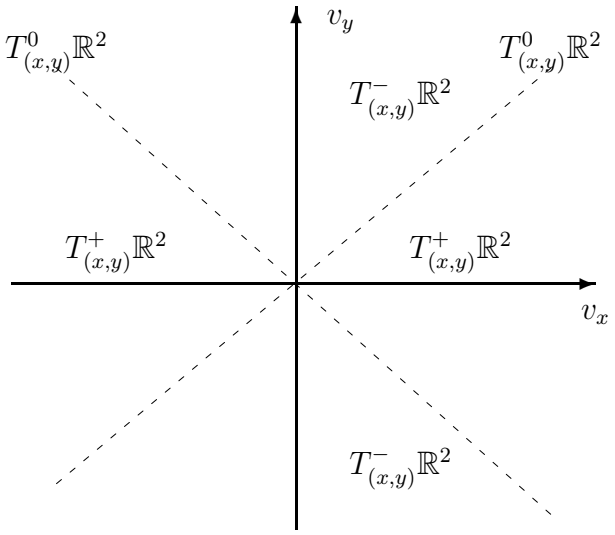

Fig. 1. Structure of the tangent space $T_{(x,y)}\mathbb{R}^2$ decomposed according to equations (76).

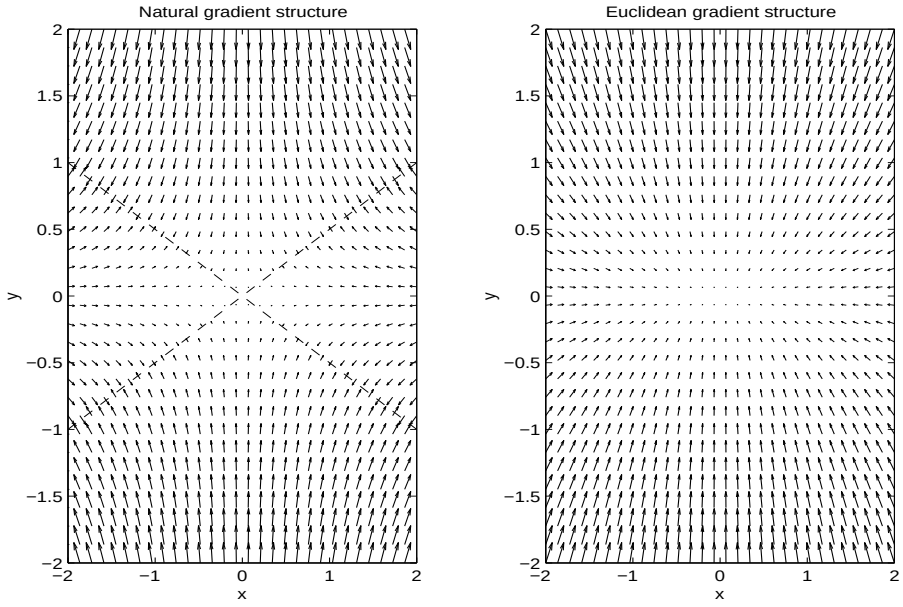

Fig. 2. Left-hand panel: Structure of the anti-natural gradient $-\nabla_{(x,y)}f(x,y)$. The dashed lines denotes the subset $T^0_{(x,y)}\mathbb{R}^2$. Right-hand panel: anti-Euclidean gradient $-(\partial_x f, \partial_y f)$.

Computing the partial derivatives of function $F_{(x,y)}$ and setting them to zero yields:

$$\frac{\partial F_{(x,y)}}{\partial v_x} = -2x + \lambda_{(x,y)} v_x = 0$$
$$\frac{\partial F_{(x,y)}}{\partial v_y} = -4y - \lambda_{(x,y)} v_y = 0,$$

hence $(v_x, v_y)\lambda_{(x,y)} = 2(x, -2y)$, therefore:

$$\nabla_{(x,y)}f(x,y) = \begin{cases} \text{Undefined}, & \text{for } x \pm 2y = 0,\ (x,y) \neq (0,0), \\ 2(x,-2y)\text{sign}(x^2 - 4y^2) & \text{otherwise.} \end{cases} \tag{78}$$

The figure 2 shows the structure of the natural gradient (78) with sign reversed in comparison to the Euclidean gradient $(\partial_x f, \partial_y f)$ (with sign reversed) of the criterion function.

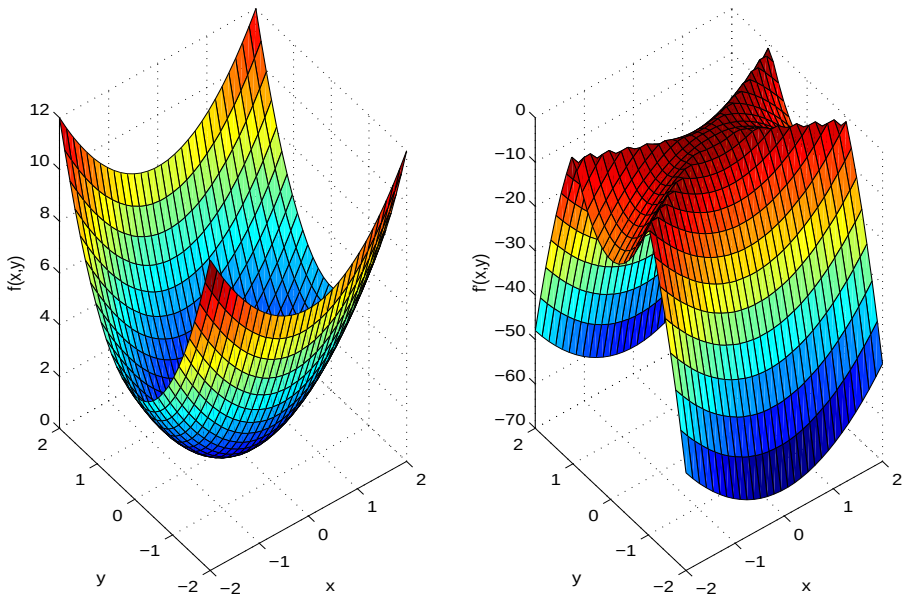

Fig. 3. Left-hand panel: Shape of function $f(x,y)$ to minimize. Right-hand panel: Rate of change of the function f at any point of the domain induced by the differential equation $(\dot{x},\dot{y}) = -\nabla_{(x,y)} f$.

The structure of the geodesic arcs on the space $(\mathbb{R}^2, \langle .,. \rangle_{(x,y)})$ may be defined via the variational principle:

$$\delta \int_0^1 \langle (\dot{x},\dot{y}),(\dot{x},\dot{y}) \rangle_{(x,y)} dt = \delta \int_0^1 (\dot{x}^2 - \dot{y}^2) dt = 0. \tag{79}$$

Computing the variations and integrating by parts yields:

$$\int_0^1 (\ddot{x}\delta x - \ddot{y}\delta y) dt = 0.$$

As the variations δx and δy are arbitrary, the geodesic equations are simply $\ddot{x} = 0$ and $\ddot{y} = 0$, hence the expression of the geodesic is:

$$G_{(x,y),(v_x,v_y)}(t) = (x + tv_x, y + tv_y). \tag{80}$$

The differential equation $(\dot{x},\dot{y}) = -\nabla_{(x,y)} f(x,y)$ is such that:

$$\frac{d}{dt} f(x(t),y(t)) = \dot{x}\partial_x f + \dot{y}\partial_y f = 4(x^2 - 4y^2)\text{sign}(4y^2 - x^2) < 0, \tag{81}$$

whenever $x \pm 2y \neq 0$.

In the present example, the search space and its tangent planes may be identified ($T_{(x,y)}\mathbb{R}^2 \cong \mathbb{R}^2$). The figure 3 shows the function $f(x,y)$ to minimize as well as the function $f'(x,y) \stackrel{\text{def}}{=} -(\partial_x f, \partial_y f)^T \nabla_{(x,y)} f$ that represents the rate of change of the function f at any point of the domain induced by the differential equation $(\dot{x},\dot{y}) = -\nabla_{(x,y)} f$. Apparently, near the lines $x \pm 2y = 0$ the rate of change vanishes, while in the other parts of the plane it is negative. This confirms that the optimization differential equation $(\dot{x},\dot{y}) = -\nabla_{(x,y)} f$ 'seeks' for the minimum of function f except that in the light-like part of the tangent space.

The figure 4 shows an example of the map $t \mapsto f(x(t),y(t))$ induced by the differential equation $(\dot{x},\dot{y}) = -\nabla_{(x,y)} f(x,y)$, integrated numerically by the method (3), and the corresponding trajectory $(x(t),y(t))$, when the initial point is chosen randomly. The figure 4 shows clearly that the vector field $-\nabla_{(x,y)} f(x,y)$ is such that it pushes the trajectory $(x(t),y(t))$ toward the singularity set $\{(x,y) \in \mathbb{R}^2 | x \pm 2y = 0\}$, where the quantity $\frac{d}{dt} f(x(t),y(t))$ gets close to zero. However, in the

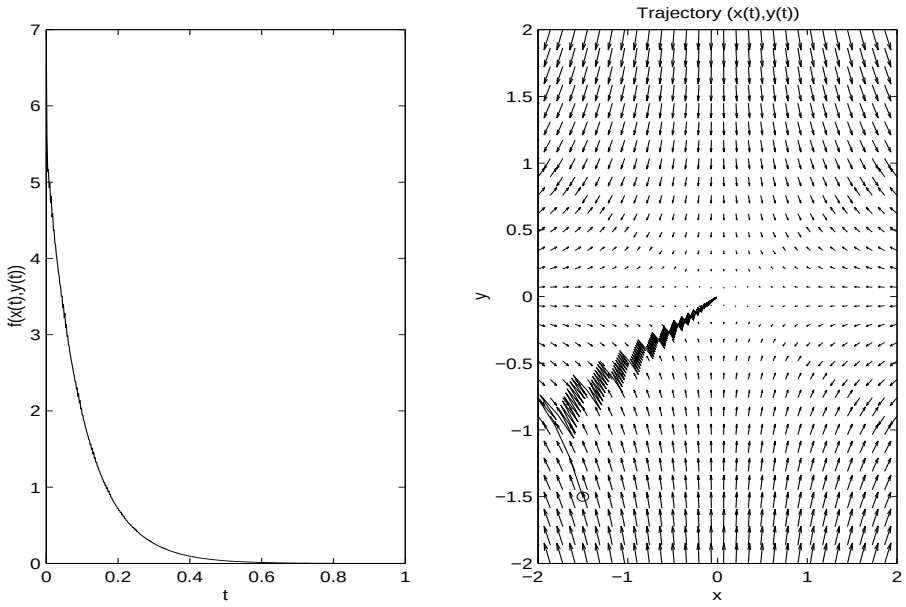

Fig. 4. Left-hand panel: Example of map $t \mapsto f(x(t), y(t))$ induced by the differential equation $(\dot{x}, \dot{y}) = -\nabla_{(x,y)} f(x, y)$. Right-hand panel: Corresponding trajectory $(x(t), y(t))$. The initial point, chosen randomly, is identified by an open circle.

proximity of the singularity 'cross' $x \pm 2y = 0$, the gradient vector field itself does not vanish (except for the point $(0, 0)$), therefore the singularity set is not stable. Accordingly, the trajectory moves toward the equilibrium point (that is the point of minimum value of the function $f(x, y)$), oscillating around one of the arms of the singularity cross.

To investigate further on the numerical behavior of the learning algorithm near singularity, it is worth considering the pathological case that the algorithm is initialized over the singularity cross. The figure 5 shows again an example of the map $t \mapsto f(x(t), y(t))$ induced by the differential equation $(\dot{x}, \dot{y}) = -\nabla_{(x,y)} f(x, y)$ integrated numerically by the method (3) and the corresponding trajectory $(x(t), y(t))$, when the initial point is chosen over the set $\{(x, y) \in \mathbb{R}^2 | x \pm 2y = 0\}$ at random. The figure 5 shows that, again, the trajectory moves toward the equilibrium point oscillating around one of the arms of the singularity cross.

It is interesting to note that, in both cases, the trajectory corresponds to a monotonically decreasing shape of the induced map $t \mapsto f(x(t), y(t))$.

B. Learning by optimization on matrix-type manifolds

Let us consider the case of minimizing the quantity $f(x) = \frac{1}{2}|d^2(g, x)|$ over the general linear group $Gl(3)$, where $g \in Gl(3)$ is randomly generated. (In the simulations, every entry of the matrix g was drawn from a normal distribution. A sanity check that $\det(g) \neq 0$ was performed.) The absolute-value operator is necessary because the squared distance $d^2(.,.)$ might assume negative values, hence the minimization of the map $x \mapsto d^2(., x)$ over the non-compact manifold $Gl(n)$ would not make sense. It is understood that the function f is not differentiable at those points for which $d(g, x) = 0$.

The figure 6 shows the course of the function $f(x(t))$ and the corresponding course of the gradient squared norm $\langle \dot{x}(t), \dot{x}(t) \rangle_{x(t)}$. Five independent trials are shown. Real symplectic matrices possess several applications. For instance, in computational ophthalmology, it is assumed that the first-order optical nature of a centered optical system is completely described by a real symplectic matrix [12]. Also, real symplectic groups play an important role in quantum mechanics [11], an important application of which is quantum computing.

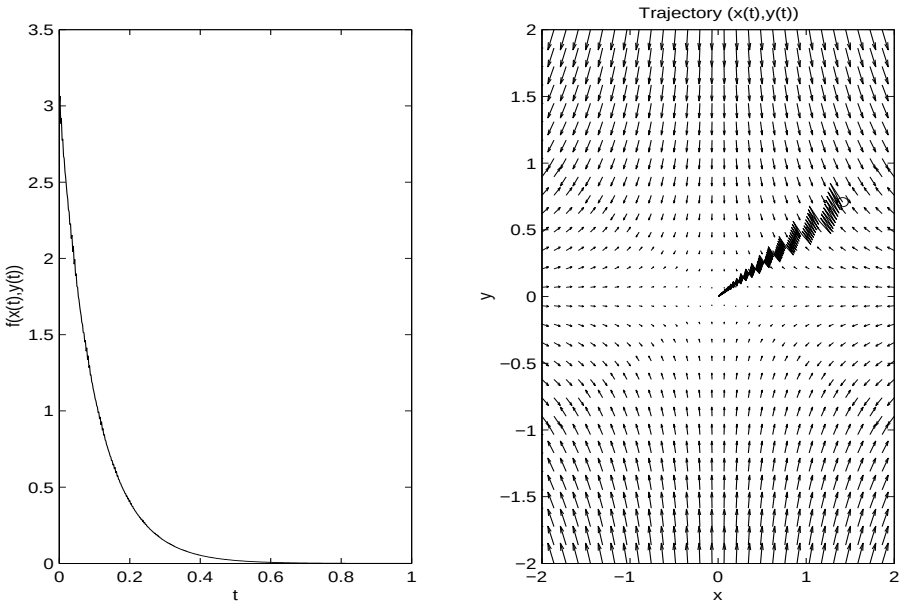

Fig. 5. Left-hand panel: Example of map $t \mapsto f(x(t), y(t))$ induced by the differential equation $(\dot{x}, \dot{y}) = -\nabla_{(x,y)} f(x, y)$. Right-hand panel: Corresponding trajectory $(x(t), y(t))$. The initial point, chosen at random over the singularity set $\{(x, y) \in \mathbb{R}^2 | x \pm 2y = 0\}$, is identified by an open circle.

In view of numerical testing, it is useful to define a way to generate random $Sp(2n)$-matrices. The notion of geodesic arc may be exploited to this purpose. Let us denote by $r \in \mathbb{R}^{2n \times 2n}$ an arbitrary random matrix and by $\tau \in [0\ 1]$ a randomly-selected real number. Set $h \stackrel{\text{def}}{=} \frac{1}{2} q(r + r^T)$. A random $Sp(2n)$-matrix close to the identity may be generated through the geodesic formula $x = \exp(\tau h)$. Analogously, a random $Sp(2n)$-matrix close to any given $Sp(2n)$-matrix x may be generated by the geodesic formula $x \exp(\tau h)$. (In the simulations, every entry of the matrix r was drawn from a zero-mean Gaussian distribution with variance $\frac{1}{100}$. The number τ was set to $\frac{1}{2}$.)

Let us consider the case of minimizing the quantity $f(x) = \frac{1}{2} |d^2(g, x)|$ over the symplectic group $Sp(4)$, where $g \in Sp(4)$ is randomly generated. It is again understood that the function f is not differentiable at those points for which $d(g, x) = 0$.

The above problem arises, e.g., when the center of mass of a cloud $\{y^{(p)}\}_{p=1,\ldots,P}$ of P points $y^{(p)} \in Sp(2n)$ is sought for. In fact, the center of mass [9], [10] may be defined on a pseudo-Riemannian manifold, as the minimizer of the function $x \mapsto \frac{1}{P} \sum_p |d^2(x, y^{(p)})|$.

The figure 7 shows the course of the function $f(x(t))$ and the corresponding course of the gradient squared norm $\langle \dot{x}(t), \dot{x}(t) \rangle_{x(t)}$. Results of ten independent trials are shown (two curves are superimposed).

In order to provide a statistical characterization of the optimization properties of the proposed geodesic-stepping-based learning algorithm, a Monte-Carlo test was performed. The optimization problem is on the symplectic group $Sp(10)$ and the Monte-Carlo test was performed on 200 independent trials. The results are show in the figure 8. The figure confirms that in every trial the algorithm reaches very low values of the criterion function f (the optimal value is 0 or $-\infty$ in a logarithmic scale that was made use of in the figure).

V. Conclusions

Although well-studied and applied to learning, Riemannian natural gradient theory is of limited impact because Riemannian geometry does not enable one to compute closed-form expressions for non-compact manifolds of interest.

The present manuscript discusses the notion of natural gradient associated to a pseudo-Riemannian metric, along with the

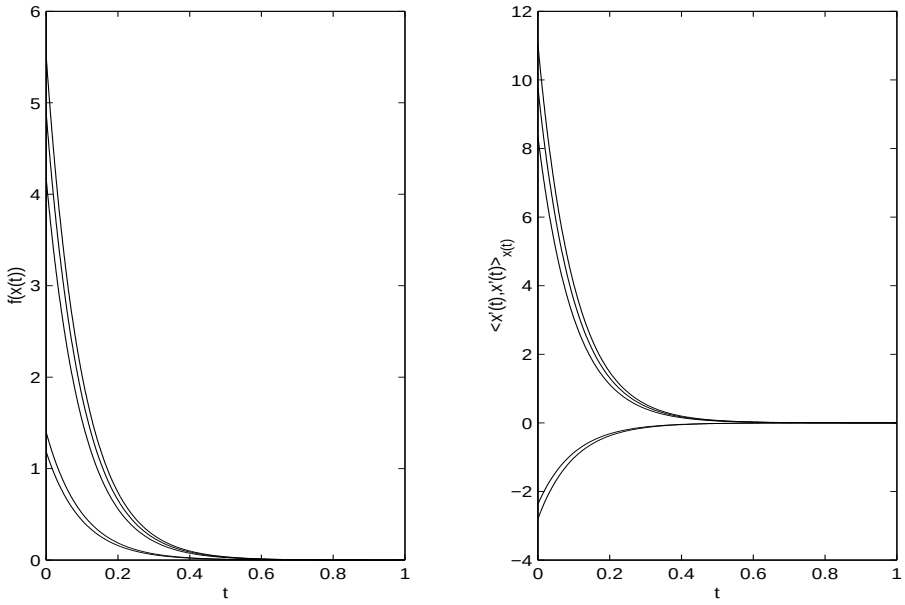

Fig. 6. Optimization of a squared-geodesic-distance criterion over the real general linear group $Gl(3)$. Left-hand panel: Course of the function $f(x(t))$. Right-hand panel: Corresponding course of the gradient squared norm $\langle \dot{x}(t), \dot{x}(t) \rangle_{x(t)}$.

companion notion of geodesics and geodesic distance on a pseudo-Riemannian manifold. In particular, a pseudo-Riemannian matric was recalled from literature and was applied to two non-compact manifolds of interest.

The aim of the manuscript is to show that endowing a non-compact manifold with a pseudo-Riemannian metric instead of a Riemannian metric might be profitable in whose cases in which pseudo-Riemannian geometry allows to calculate the closed form of entities of interest, such as geodesic arcs and geodesic distances.

The related learning theory, based on pseudo-Riemannian natural gradient and pseudo-Riemannian geodesic stepping, is indeed more general and may be applied to general abstract manifolds.

VI. Acknowledgments

I wish to gratefully thank S.-i. Amari, E. Celledoni and Y. Nishimori for their fruitful comments and suggestions on a earlier version of the present paper.

References

[1] P.-A. Absil and K.A. Gallivan, *Joint diagonalization on the oblique manifold for independent component analysis*, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06), Vol. 5, pp. pp. 945 – 948, 2006

[2] P.-A. Absil, R. Mahony and R. Sepulchre, *Riemannian geometry of Grassmann manifolds with a view on algorithmic computation*, Acta Applicandae Mathematicae, Vol. 80, No. 2, pp. 199 – 220, January 2004

[3] B. Afsari and P.S. Krishnaprasad, *Some gradient based joint diagonalization methods for ICA*, Proceedings of the 5^{th} International Conference on Independent Component Analysis and Blind Source Separation (Springer LCNS Series), 2004

[4] S.-i. Amari, *Natural gradient works efficiently in learning*, Neural Computation, Vol. 10, 251 – 276, 1998

[5] A.M. Bloch, P.E. Crouch, J.E. Marsden and A.K. Sayal, *Optimal control and geodesics on quadratic matrix Lie groups*, Foundations of Computational Mathematics, Vol. 8, pp. 469 – 500, 2008

[6] S. Fiori, *Quasi-geodesic neural learning algorithms over the orthogonal group: A tutorial*, Journal of Machine Learning Research, Vol. 6, pp. 743 – 781, May 2005

[7] S. Fiori, *Lie-group-type neural system learning by manifold retractions*, Neural Networks (Elsevier), Vol. 21, No. 10, pp. 1524 – 1529, December 2008

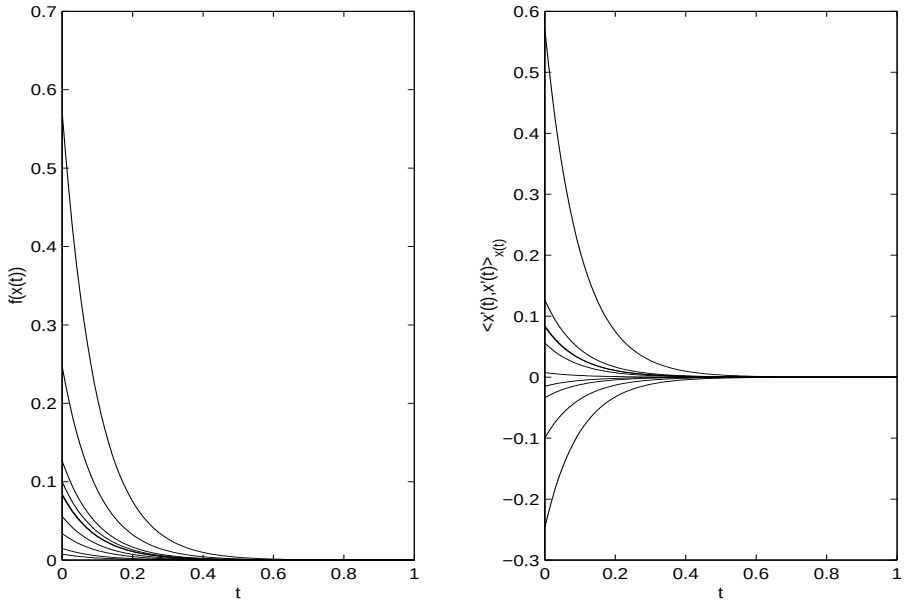

Fig. 7. Optimization of a squared-geodesic-distance criterion over the real symplectic group $Sp(4)$. Left-hand panel: Course of the function $f(x(t))$. Right-hand panel: Corresponding course of the gradient squared norm $\langle \dot{x}(t), \dot{x}(t) \rangle_{x(t)}$.

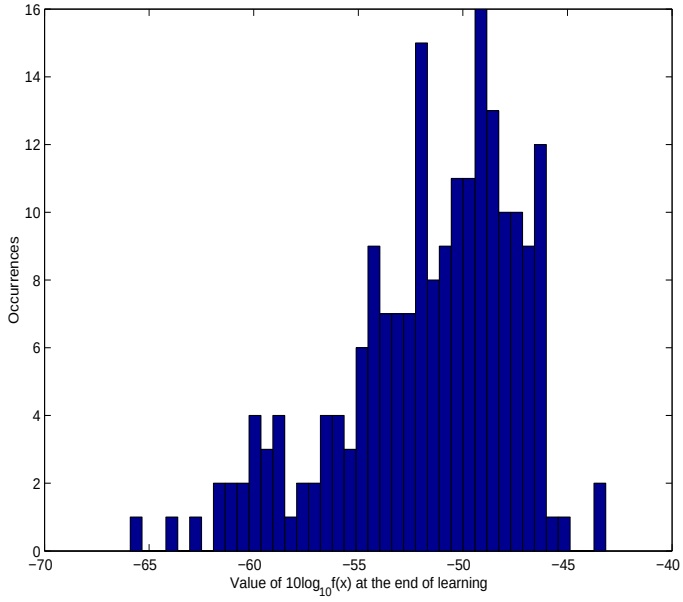

Fig. 8. Optimization of a squared-geodesic-distance criterion over the real symplectic group $Sp(10)$. Learning criterion after learning.

[8] S. FIORI, *Learning the Fréchet mean over the manifold of symmetric positive-definite matrices*, Cognitive Computation (Springer), Vol. 1, No. 4, pp. 279 – 291, December 2009

[9] S. FIORI AND T. TANAKA, *An algorithm to compute averages on matrix Lie groups*, IEEE Trans. on Signal Processing, Vol. 57, No. 12, pp. 4734 – 4743, December 2009

[10] M. FRÉCHET, *Les élements aléatoires de nature quelconque dans un espace distancié*, Annales de l'Institut Henri Poincaré, Vol. 10, pp 215 – 310, 1948

[11] V. GUILLEMIN AND S. STERNBERG, *Symplectic Techniques in Physics*, Cambridge University Press, 1984

[12] W.F. HARRIS, *Paraxial ray tracing through noncoaxial astigmatic optical systems, and a* 5×5 *augmented system matrix*, Optometry and Vision Science, Vol. 71, No. 4, pp. 282 – 285, 1994

[13] D.G. LUENBERGER, *The gradient projection method along geodesics*, Management Science, Vol. 18, No. 11, pp. 620 – 631, July 1972

[14] B. KHESIN AND S. TABACHNIKOV, *Pseudo-Riemannian geodesics and billiards*, Advances in Mathematics, Vol. 221, No. 4, pp. 1364 – 1396, July 2009

[15] A. KHVEDELIDZE AND D. MLADENOV, *Generalized Calogero-Moser-Sutherland models from geodesic motion on $Gl^{+}(n,\mathbb{R})$ group manifold*, Physics Letters A, Vol. 299, No.s 5-6, pp. 522 – 530, July 2002

[16] A. PAMBIRA, *Harmonic morphisms between degenerate semi-Riemannian manifolds*, Beiträge zur Algebra und Geometrie, Contributions to Algebra and Geometry, Vol. 46, No. 1, pp. 261 – 281, 2005

[17] C. SMALL, *The Statistical Theory of Shape*. Springer, 1996

[18] M. SPIVAK, *A Comprehensive Introduction to Differential Geometry*, 2^{nd} Edition, Berkeley, CA: Publish or Perish Press, 1979

[19] A. YEREDOR, *Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation*, IEEE Trans. on Signal Processing, vol. 50, No. 7, pp. 1545 – 1553, July 2002